

Joint Probabilistic Linear Discriminant Analysis

Luciana Ferrer¹

¹Instituto de Investigación en Ciencias de la Computación,
CONICET-UBA, Argentina

Abstract

Standard probabilistic discriminant analysis (PLDA) for speaker recognition assumes that the sample's features (usually, i-vectors) are given by a sum of three terms: a term that depends on the speaker identity, a term that models the within-speaker variability and is assumed independent across samples, and a final term that models any remaining variability and is also independent across samples. In this work, we propose a generalization of this model where the within-speaker variability is not necessarily assumed independent across samples but dependent on another discrete variable. This variable, which we call the channel variable as in the standard PLDA approach, could be, for example, a discrete category for the channel characteristics, the language spoken by the speaker, the type of speech in the sample (conversational, monologue, read), etc. The value of this variable is assumed to be known during training but not during testing. Scoring is performed, as in standard PLDA, by computing a likelihood ratio between the null hypothesis that the two sides of a trial belong to the same speaker versus the alternative hypothesis that the two sides belong to different speakers. The two likelihoods are computed by marginalizing over two hypothesis about the channels in both sides of a trial: that they are the same and that they are different. This way, we expect that the new model will be better at coping with same-channel versus different-channel trials than standard PLDA, since knowledge about the channel (or language, or speech style) is used during training and implicitly considered during scoring.

1 Introduction

PLDA [1] was first proposed for doing inferences about the identity of a person from an image of their face. The technique was later widely adopted by the speaker recognition community, becoming the state-of-the-art scoring technique for this task. In this work, we will adopt the nomenclature usually used in the speaker recognition community. Yet, the model proposed can be used for the original image processing task or any other task for which standard PLDA is used.

Standard PLDA assumes that the vector m_i representing a certain sample (in speaker recognition these will usually be i-vectors [2]) from speaker s_i is given by

$$m_i = Vy_{s_i} + Ux_i + z_i \quad (1)$$

where y_{s_i} is a vector of size R_y (the dimension of the speaker subspace) and x_i is a vector of size R_x (the dimension of the channel subspace), and

$$y_{s_i} \sim N(0, I) \quad (2)$$

$$x_i \sim N(0, I) \quad (3)$$

$$z_i \sim N(0, D^{-1}) \quad (4)$$

where the matrix D is assumed to be diagonal. All these latent variables are assumed independent: speaker variables are independent across speakers and the channel variable x_i and noise variable z_i are independent across samples.

The model described corresponds to the original PLDA formulation. In speaker recognition a simplified version of PLDA is more commonly used, where the matrix V is full rank and the

channel factor is absorbed into the noise factor, which is then assumed to have a full rather than diagonal covariance matrix. This simpler model has shown to give better performance than the original model. See [3] for a comprehensive explanation of the usual flavors of PLDA.

In this work, we propose a generalization of the original model where the channel variable is no longer considered independent across samples, but potentially shared (tied) across samples from different speakers.

2 Proposed Joint PLDA Model

The proposed generalization of the PLDA model implies that the channel latent variable is no longer dependent only on the sample index, but rather, depends on a separate channel label. This makes the model symmetric in the two latent variables (speaker and channel, in our nomenclature) in the sense that both variables are tied across all samples sharing a certain label. To represent this dependency, we introduce a channel label for each sample, called c_i . Given this channel label, and the speaker label s_i , we propose to model vector m_i for sample i as:

$$m_i = Vy_{s_i} + Ux_{c_i} + z_i \quad (5)$$

where, as before, y_{s_i} is a vector of size R_y and x_{c_i} is a vector of size R_x , and

$$y_{s_i} \sim N(0, I) \quad (6)$$

$$x_{c_i} \sim N(0, I) \quad (7)$$

$$z_i \sim N(0, D^{-1}) \quad (8)$$

The model's parameters to estimate are $\lambda = \{V, U, D\}$, as in the standard PLDA formulation, but the input data for the training algorithm is now expected to have a second set of labels indicating the channel identity of each sample. Note that while we call this variable the channel, for consistency with previous work on PLDA, the channel could be anything that is a nuisance variable for the task. For speaker recognition this could be, for example, a discrete classification of the channel type itself, or the language spoken, the speech style, etc.

The following sections derive the probabilities that are needed during training with the expectation-maximization algorithm and during scoring with the likelihood ratio. The derivations closely follow the ones for standard PLDA in [4] with one main difference: in the new model, most probabilities cannot be formulated by speaker and then multiplied to get the total probabilities, as is usually done for standard PLDA, since the channel introduces dependencies across samples from different speakers. Instead, we formulate all probabilities over all samples.

In the following, we take

$$Y = \{y_1, \dots, y_S\} \quad (9)$$

$$X = \{x_1, \dots, x_C\} \quad (10)$$

where S is the total number of speakers and C is the total number of channels.

2.1 Prior

The joint prior for the hidden variables for all the data is given by

$$p(Y, X) = p(X)p(Y) \propto \exp\left(-\frac{1}{2} \sum_s y_s^T y_s - \frac{1}{2} \sum_c x_c^T x_c\right) \quad (11)$$

2.2 Likelihood

The full data likelihood is given by

$$\begin{aligned}
p(M|Y, X, \lambda) &= \prod_i N(m_i | Vy_{s_i} + Ux_{c_i}, D^{-1}) \\
&\propto \exp \sum_i \left(-\frac{1}{2} (m_i - Vy_{s_i} - Ux_{c_i})^T D (m_i - Vy_{s_i} - Ux_{c_i}) + \frac{1}{2} \log |D| \right) \\
&= \exp \sum_i \left(-\frac{1}{2} m_i^T D m_i + m_i^T D V y_{s_i} + m_i^T D U x_{c_i} \right. \\
&\quad \left. - \frac{1}{2} y_{s_i}^T V^T D V y_{s_i} - y_{s_i}^T V^T D U x_{c_i} - \frac{1}{2} x_{c_i}^T U^T D U x_{c_i} + \frac{1}{2} \log |D| \right) \quad (12)
\end{aligned}$$

2.3 Joint

The joint probability is given by the product of the likelihood and the prior,

$$\begin{aligned}
p(M, Y, X | \lambda) &\propto \exp \left[\sum_i \left(-\frac{1}{2} m_i^T D m_i + m_i^T D V y_s + m_i^T D U x_c \right. \right. \\
&\quad \left. \left. - \frac{1}{2} y_{s_i}^T V^T D V y_{s_i} - y_{s_i}^T V^T D U x_{c_i} - \frac{1}{2} x_{c_i}^T U^T D U x_{c_i} \right) \right. \\
&\quad \left. - \frac{1}{2} \sum_s y_s^T y_s - \frac{1}{2} \sum_c x_c^T x_c \right] \\
&= \exp \left[\sum_i \left(-\frac{1}{2} m_i^T D m_i + m_i^T D V y_{s_i} + m_i^T D U x_{c_i} - x_{c_i}^T J y_{s_i} \right) \right. \\
&\quad \left. - \frac{1}{2} \sum_s y_s^T L_s y_s - \frac{1}{2} \sum_c x_c^T K_c x_c \right] \quad (13)
\end{aligned}$$

where

$$J = U^T D V \quad (14)$$

$$K_c = n_c U^T D U + I \quad (15)$$

$$L_s = n_s V^T D V + I \quad (16)$$

where n_c is the number of samples for channel c and n_s is the number of samples for speaker s .

2.4 Posterior

We compute the posterior from two factors:

$$p(Y, X | M, \lambda) = p(Y | X, M, \lambda) p(X | M, \lambda) \quad (17)$$

2.4.1 Outer posterior

The outer posterior is proportional (as a function of Y) to the joint probability. Keeping only the terms in the joint that depend on Y we get

$$\begin{aligned}
p(Y|X, M, \lambda) &\propto p(M, X, Y|\lambda) \\
&\propto \exp \left[\sum_i (m_i^T DV y_{s_i} - x_{c_i}^T J y_{s_i}) - \frac{1}{2} \sum_s y_s^T L_s y_s \right] \\
&\propto \exp \left[\sum_i (m_i^T DV - x_{c_i}^T J) y_{s_i} - \frac{1}{2} \sum_s y_s^T L_s y_s \right] \\
&\propto \exp \sum_s \left[y_s^T \sum_{i|s_i=s} (V^T D m_i - J^T x_{c_i}) - \frac{1}{2} y_s^T L_s y_s \right] \\
&\propto \exp \sum_s \left[y_s^T (V^T D f_s - J^T \bar{x}_s) - \frac{1}{2} y_s^T L_s y_s \right] \\
&\propto \prod_s N(y_s | \hat{y}_s, L_s^{-1})
\end{aligned} \tag{18}$$

where

$$f_s = \sum_{i|s_i=s} m_i \tag{19}$$

$$\bar{x}_s = \sum_{i|s_i=s} x_{c_i} \tag{20}$$

$$\tilde{y}_s = L_s^{-1} V^T D f_s \tag{21}$$

$$\hat{y}_s = \tilde{y}_s - L_s^{-1} J^T \bar{x}_s \tag{22}$$

2.4.2 Inner posterior

To get the posterior, which is proportional to the joint between X and M , we use a nice trick:

$$\begin{aligned}
p(X|M, \lambda) &\propto p(M, X|\lambda) = \frac{p(Y, X, M|\lambda)}{p(Y|X, M, \lambda)} \Big|_{Y=0} \\
&\propto \frac{\exp(\sum_i m_i^T D U x_{c_i} - \frac{1}{2} \sum_c x_c^T K_c x_c)}{\exp(-\frac{1}{2} \sum_s \hat{y}_s^T L_s \hat{y}_s)}
\end{aligned} \tag{23}$$

where the right hand side in the first line is independent of Y (since the left-hand side is) and, hence, can be conveniently evaluated at 0 ([5]). Now, we expand the exponent in the denominator

$$\begin{aligned}
\sum_s \hat{y}_s^T L_s \hat{y}_s &= \sum_s (\tilde{y}_s^T - \bar{x}_s^T J L_s^{-1}) (L_s \tilde{y}_s - J^T \bar{x}_s) \\
&= \sum_s (-2 \bar{x}_s^T J \tilde{y}_s + \bar{x}_s^T J L_s^{-1} J^T \bar{x}_s) + \text{const}
\end{aligned} \tag{24}$$

and plug it back in the posterior:

$$\begin{aligned}
p(X|M, \lambda) &\propto \exp \left(\sum_i m_i^T D U x_{c_i} - \frac{1}{2} \sum_c x_c^T K_c x_c - \sum_s \bar{x}_s^T J \tilde{y}_s + \frac{1}{2} \sum_s \bar{x}_s^T J L_s^{-1} J^T \bar{x}_s \right) \\
&= \exp \left(\sum_c g_c^T D U x_c - \frac{1}{2} \sum_c x_c^T K_c x_c - \sum_i x_{c_i}^T J \tilde{y}_{s_i} + \frac{1}{2} \sum_s \bar{x}_s^T J L_s^{-1} J^T \bar{x}_s \right) \\
&= \exp \left(\sum_c g_c^T D U x_c - \frac{1}{2} \sum_c x_c^T K_c x_c - \sum_c x_c^T J \tilde{y}_c + \frac{1}{2} \sum_s \bar{x}_s^T J L_s^{-1} J^T \bar{x}_s \right)
\end{aligned} \tag{25}$$

where

$$g_c = \sum_{i|c_i=c} m_i \quad (26)$$

$$\bar{\tilde{y}}_c = \sum_{i|c_i=c} \tilde{y}_{s_i} = \sum_{i|c_i=c} L_{s_i}^{-1} V^T D f_{s_i} \quad (27)$$

Since there is no way to disentangle the x s for different c s, we need to obtain the distribution for all x s at once. We define vectors which are the concatenation of all individual vectors:

$$\mathbf{X} = [x_1^T \dots x_C^T]^T \quad (28)$$

$$\mathbf{Y} = [y_1^T \dots y_S^T]^T \quad (29)$$

and, similarly, for all other vectors. Converting the sums into matrix form, we get

$$\begin{aligned} p(X|M, \lambda) &\propto \exp \left(\mathbf{X}^T M_1 \mathbf{G} - \frac{1}{2} \mathbf{X} M_2 \mathbf{X} - \mathbf{X} M_3 \bar{\tilde{\mathbf{Y}}} + \frac{1}{2} \bar{\tilde{\mathbf{X}}}^T M_4 \bar{\tilde{\mathbf{X}}} \right) \\ &= \exp \left(\mathbf{X}^T M_1 \mathbf{G} - \frac{1}{2} \mathbf{X} M_2 \mathbf{X} - \mathbf{X} M_3 \bar{\tilde{\mathbf{Y}}} + \frac{1}{2} \mathbf{X}^T H^T M_4 H \mathbf{X} \right) \\ &= \exp \left(\mathbf{X}^T (M_1 \mathbf{G} - M_3 \bar{\tilde{\mathbf{Y}}}) - \frac{1}{2} \mathbf{X} (M_2 - H^T M_4 H) \mathbf{X} \right) \end{aligned} \quad (30)$$

where

$$M_1 = \text{diag}(U^T D, C) \quad (31)$$

$$M_2 = \text{diag}(K_1, \dots, K_C) \quad (32)$$

$$M_3 = \text{diag}(J, S) \quad (33)$$

$$M_4 = M_3 \text{diag}(L_1^{-1}, \dots, L_S^{-1}) M_3^T \quad (34)$$

$$\bar{\tilde{\mathbf{X}}} = H \mathbf{X} \quad (35)$$

where $\text{diag}(M, N)$ is a block diagonal matrix with matrix M in each of N blocks and $\text{diag}(M_1, \dots, M_N)$ is a block diagonal matrix with blocks given by matrices M_i . The matrix H is of size $S R_x \times C R_x$, where block $H_{s,c}$ (R_x rows and columns starting at position $(s R_x, c R_x)$ in H) is given by:

$$H_{s,c} = n_{s,c} I \quad (36)$$

where I is the identity matrix of size R_x and $n_{s,c}$ is the number of times that channel s occurs for speaker s , which could be zero.

Hence, finally:

$$p(\mathbf{X}|M, \lambda) = N(\mathbf{X}|\hat{\mathbf{X}}, \Sigma) \quad (37)$$

where

$$\Sigma = (M_2 - H^T M_4 H)^{-1} \quad (38)$$

$$\hat{\mathbf{X}} = \Sigma (M_1 \mathbf{G} - M_3 \bar{\tilde{\mathbf{Y}}}) \quad (39)$$

Now if we need the distribution of x_c we just need to get the marginal distribution from the one above, which means getting the c block from the mean and covariance matrix. So,

$$p(x_c|M, \lambda) = N(x_c|\hat{x}_c, \Sigma_c) \quad (40)$$

$$\Sigma_c = \text{block}(\Sigma, c, c) \quad (41)$$

$$\hat{x}_c = \text{block}(\hat{\mathbf{X}}, c) \quad (42)$$

3 EM algorithm

As in standard PLDA, we will use the expectation-maximization algorithm to train the model parameters.

3.1 EM objective

The objective function of EM is the likelihood of the data given the model, which can be obtained as follows:

$$\begin{aligned}
\log p(M|\lambda) &= \log \frac{p(M|Y, X, \lambda)p(X)p(Y)}{p(Y|X, M, \lambda)p(X|M, \lambda)} \Big|_{Y=0, X=0} \\
&= \sum_i \left(-\frac{1}{2} m_i^T D m_i + \frac{1}{2} \log |D| \right) - \sum_s \left(-\frac{1}{2} \tilde{y}_s^T L_s \tilde{y}_s + \frac{1}{2} \log |L_s| \right) \\
&\quad + \frac{1}{2} \hat{\mathbf{X}}^T \Sigma^{-1} \hat{\mathbf{X}} + \frac{1}{2} \log |\Sigma| + \text{constant}
\end{aligned} \tag{43}$$

Terms that depend only on data or are constant have been discarded since they will not change with model's parameters.

3.2 EM auxiliary function

The EM auxiliary function is given by the expected value of the log-likelihood with respect to the posterior probability of the hidden variables given the data and the previously estimated model parameters, λ_{k-1} . Defining $Z = \{X, Y\}$, then

$$\begin{aligned}
Q(\lambda_k | \lambda_{k-1}) &= E_{Z|M, \lambda_{k-1}} [\log p(M, Z | \lambda_k)] \\
&= \langle \log p(M|Z, \lambda_k) p(Z | \lambda_k) \rangle \\
&= \langle \log p(M|Z, \lambda_k) \rangle + \text{const} \\
&= \sum_i \langle \frac{1}{2} \log |D| - \frac{1}{2} (m_i - W z_i)^T D (m_i - W z_i) \rangle + \text{const} \\
&= \sum_i \langle \frac{1}{2} \log |D| - \frac{1}{2} m_i^T D m_i - \frac{1}{2} z_i^T W^T D W z_i + m_i^T D W z_i \rangle + \text{const} \\
&= \frac{N}{2} \log |D| - \frac{1}{2} \sum_i \langle m_i^T D m_i - \frac{1}{2} z_i^T W^T D W z_i + m_i^T D W z_i \rangle + \text{const} \\
&= \frac{N}{2} \log |D| - \frac{1}{2} \text{tr}(SD) - \frac{1}{2} \text{tr}(RW^T D W) + \text{tr}(T D W) + \text{const}
\end{aligned} \tag{44}$$

where the \langle and \rangle symbols stand for the expectation with respect to the distribution of Z given the data and the previous parameters, as in the first line, and

$$z_i = [x_{c_i}^T y_{s_i}^T]^T \tag{45}$$

$$W = [UV] \tag{46}$$

$$S = \sum_i m_i m_i^T \tag{47}$$

$$R = \sum_i \langle z_i z_i^T \rangle \tag{48}$$

$$T = \sum_i \langle z_i \rangle m_i^T \tag{49}$$

In this derivation we use the fact that $p(Z|\lambda_k)$ is a constant with respect to λ_k , since the prior for the latent variables does not depend on model's parameters (Equation (11)). We also use the fact that $m_i^T D m_i$ is a scalar and, hence, $m_i^T D m_i = \text{tr}(m_i^T D m_i) = \text{tr}(m_i m_i^T D)$ since $\text{tr}(AB) = \text{tr}(BA)$. A similar thing is done for the R term and the T term.

3.3 M-Step

Now, differentiating Q with respect to D and W and setting to zero, we get that

$$D^{-1} = \frac{1}{N}(S - WT) \quad (50)$$

$$W^T = R^{-1}T \quad (51)$$

So, the matrices are estimated exactly the same way as in the standard PLDA approach [4]. The complexity lies in getting R and T .

3.4 E-Step

To find T and R we decompose it into their x and y blocks and then use the distributions we found in Section 2.4.1. This way, $\langle x_c \rangle = \hat{x}_c$, where the right hand side is given by Equation (42), and $\langle x_c x_c^T \rangle = \Sigma_c + \hat{x}_c \hat{x}_c^T$, where Σ_c is given by Equation (41).

For the expectation of y we need to use the following the law of total expectation:

$$\begin{aligned} \langle y_s \rangle &= E_{Y|M, \lambda_{k-1}} [y_s] \\ &= E_{X|M, \lambda_{k-1}} [E_{Y|M, X, \lambda_{k-1}} [y_s | X]] \\ &= E_{X|M, \lambda_{k-1}} [L_s^{-1} V^T D f_s - L_s^{-1} J^T \bar{x}_s] \\ &= L_s^{-1} V^T D f_s - L_s^{-1} J^T E_{X|M, \lambda_{k-1}} \left[\sum_{i|s_i=s} x_{c_i} \right] \\ &= L_s^{-1} V^T D f_s - L_s^{-1} J^T \sum_{i|s_i=s} \hat{x}_{c_i} \end{aligned} \quad (52)$$

where we use Equation (22) to get the inner expectation in the second line. Similar procedures can be used to get the second moments needed to compute R .

Using all the equalities above,

$$\begin{aligned} T_x &= \sum_i \langle x_{c_i} \rangle m_i^T \\ &= \sum_c \hat{x}_c \sum_{i|c_i=c} m_i^T \\ &= \sum_c \hat{x}_c g_c^T \end{aligned} \quad (53)$$

$$\begin{aligned} T_y &= \sum_i \langle y_{s_i} \rangle m_i^T \\ &= \sum_s \hat{y}_s \sum_{i|s_i=s} m_i^T \\ &= \sum_s (L_s^{-1} V^T D f_s - L_s^{-1} J^T \sum_{i|s_i=s} \hat{x}_{c_i}) f_s^T \\ &= \sum_s L_s^{-1} V^T D f_s f_s^T - \sum_s L_s^{-1} J^T \sum_{i|s_i=s} \hat{x}_{c_i} f_s^T \\ &= \sum_s L_s^{-1} V^T D f_s f_s^T - \sum_s L_s^{-1} J^T \hat{\bar{x}}_s f_s^T \end{aligned} \quad (54)$$

where we define

$$\hat{\bar{x}}_s = \sum_{i|s_i=s} \hat{x}_{c_i} \quad (55)$$

$$\begin{aligned}
R_{xx} &= \sum_i \langle x_{c_i} x_{c_i}^T \rangle \\
&= \sum_c n_c \langle x_c x_c^T \rangle \\
&= \sum_c n_c (\Sigma_c + \hat{x}_c \hat{x}_c^T)
\end{aligned} \tag{56}$$

$$\begin{aligned}
R_{yx} &= \sum_i \langle y_{s_i} x_{c_i}^T \rangle \\
&= \sum_i \langle (L_{s_i}^{-1} V^T D f_{s_i} - L_{s_i}^{-1} J^T \sum_{j|s_j=s_i} x_{c_j}) x_{c_i}^T \rangle \\
&= \sum_i L_{s_i}^{-1} \left[V^T D f_{s_i} \hat{x}_{c_i}^T - J^T \sum_{j|s_j=s_i} \langle x_{c_j} x_{c_i}^T \rangle \right] \\
&= \sum_s L_s^{-1} \left[V^T D f_s \sum_{i|s_i=s} \hat{x}_{c_i}^T - J^T \sum_{i|s_i=s} \sum_{j|s_j=s} \langle x_{c_i} x_{c_j}^T \rangle \right] \\
&= \sum_s L_s^{-1} [V^T D f_s \bar{x}_s^T - J^T \langle \bar{x}_s \bar{x}_s^T \rangle]
\end{aligned} \tag{57}$$

$$\begin{aligned}
R_{yy} &= \sum_i \langle y_{s_i} y_{s_i}^T \rangle \\
&= \sum_s n_s \langle y_s y_s^T \rangle \\
&= \sum_s n_s [L_s^{-1} + \langle \hat{y}_s(\mathbf{X}) \hat{y}_s^T(\mathbf{X}) \rangle]
\end{aligned} \tag{58}$$

where

$$\begin{aligned}
\langle \hat{y}_s(\mathbf{X}) \hat{y}_s^T(\mathbf{X}) \rangle &= E[L_s^{-1} (V^T D f_s - J^T \bar{x}_s) (f_s^T D V - \bar{x}_s^T J) L_s^{-1}] \\
&= L_s^{-1} (V^T D f_s f_s^T D V - V^T D f_s \hat{x}_s^T J - J \hat{x}_s f_s^T D V + J^T \langle \bar{x}_s \bar{x}_s^T \rangle J) L_s^{-1}
\end{aligned} \tag{59}$$

and

$$\begin{aligned}
\langle \bar{x}_s \bar{x}_s^T \rangle &= \sum_{i|s_i=s} \sum_{j|s_j=s} \langle x_{c_j} x_{c_i}^T \rangle \\
&= \sum_{i|s_i=s} \sum_{j|s_j=s} [\hat{x}_{c_i} \hat{x}_{c_j}^T + \text{block}(\Sigma, c_j, c_i)]
\end{aligned} \tag{60}$$

4 Scoring

In scoring, given two sets of i-vectors E and T for enrollment and test, we need to compute:

$$LR = \frac{p(E, T | H_{SS})}{p(E, T | H_{DS})} \tag{61}$$

where SS stands for same speaker and DS for different speaker. In standard PLDA, the denominator can be factorized as $p(E)p(T)$ but this is only because it is assumed that channel factors are independent across samples. Since we do not assume that, we need to compute LR as follows, where we assume E and T are single i-vectors rather than sets.

$$LR = \frac{p(E, T | H_{SS}, H_{SC}) P(H_{SC} | H_{SS}) + p(E, T | H_{SS}, H_{DC}) P(H_{DC} | H_{SS})}{p(E, T | H_{DS}, H_{SC}) P(H_{SC} | H_{DS}) + p(E, T | H_{DS}, H_{DC}) P(H_{DC} | H_{DS})} \tag{62}$$

Here, SC stands for same channel and DC for different channel. The priors for these two hypothesis given the DS and SS hypothesis would be task-dependent (i.e., similarly to the same-gender and different-gender priors for gender-based mixture PLDA [6]).

Note that the formulation would become more complex if there was more than one sample allowed in testing or enrollment, since there could be any combination of channels for those samples, some of them being the same, some different. For now, we will focus on the case in which both E and T are single i-vectors $\{m_E\}$ and $\{m_T\}$.

Define $M = \{E, T\}$ and the latent variables for those two i-vectors to be X and Y . We can now write:

$$p(M|H_{*S}, H_{*C}) = \frac{p(M|X_{H_{*C}}, Y_{H_{*S}})p(X_{H_{*C}})(Y_{H_{*S}})}{p(X_{H_{*C}}, Y_{H_{*S}}|M)} \Big|_{X_{H_{*C}}=0, Y_{H_{*S}}=0} \quad (63)$$

where, on the right hand side, we drop the conditioning to the hypotheses to simplify notation, and where

$$X_{H_{*C}} = \begin{cases} \{x\}, & \text{if } H_{*C} = H_{SC} \\ \{x_E, x_T\}, & \text{if } H_{*C} = H_{DC} \end{cases} \quad (64)$$

and

$$Y_{H_{*S}} = \begin{cases} \{y\}, & \text{if } H_{*S} = H_{SS} \\ \{y_E, y_T\}, & \text{if } H_{*S} = H_{DS} \end{cases} \quad (65)$$

Now, the likelihood in the numerator will be the same for all four cases since regardless of whether the latent variables are tied or not, Equation (12) has the same form. Hence, that term cancels out in the computation of the LR. The priors, on the other hand, will have one factor for the same-speaker or same-channel case and two identical factors, once evaluated at 0, for the different-speaker or different-channel case. So,

$$LR = \frac{p(x)p(y)p(X_{H_{SC}}, Y_{H_{SS}}|M)^{-1}P_{S,S} + p(x)^2p(y)p(X_{H_{DC}}, Y_{H_{SS}}|M)^{-1}P_{D,S}}{p(x)p(y)^2p(X_{H_{SC}}, Y_{H_{DS}}|M)^{-1}P_{S,D} + p(x)^2p(y)^2p(X_{H_{DC}}, Y_{H_{DS}}|M)^{-1}P_{D,D}} \Big|_0 \quad (66)$$

where we have shortened the names for the hypothesis priors: $P_{S,S} = P(H_{SC}|H_{SS})$, $P_{D,S} = P(H_{DC}|H_{SS})$, and so on. The evaluation at zero is done for all X s and Y s.

All that is left to do is compute the posteriors, which are given by the product of the inner and outer posteriors, and evaluate them at 0.

$$p(X_{H_{*C}}, Y_{H_{*S}}|M) = p(Y_{H_{*S}}|X_{H_{*C}}, M)p(X_{H_{*C}}|M) \quad (67)$$

The outer posterior $p(Y_{H_{*S}}|X_{H_{*C}}, M)$ evaluated at 0 will be the same regardless of whether the channels are the same or different.

$$p(Y_{H_{*S}}|X_{H_{*C}}, M)|_0 = \begin{cases} N(y=0|L_S^{-1}V^TD(m_E+m_T), L_S^{-1}), & \text{if } H_S = H_{SS} \\ N(y=0|L_D^{-1}V^TDm_E, L_D^{-1})N(y=0|L_D^{-1}V^TDm_T, L_D^{-1}), & \text{if } H_S = H_{DS} \end{cases} \quad (68)$$

where $L_D = V^TDV + I$ and $L_S = 2V^TDV + I$.

The inner posterior is given in Equation (37), with $\mathbf{X} = x$ for the same-channel case and $\mathbf{X} = [x_E^T x_T^T]^T$ for the different-channel case. These posteriors also depend on whether the speakers are the same or not (through the implicit conditioning on the hypothesis that we dropped in (63)), which comes into play when evaluating Equations (33), (34) and (35).

5 Conclusions

We have proposed a generalization of PLDA for speaker recognition where channel factors are no longer considered independent across samples. This paper derives the formulae necessary to train the model through the expectation-maximization algorithm and to compute likelihood ratios for scoring.

The proposed method can be used for any task for which standard PLDA is used, when a discrete nuisance factor is known during training. Examples include multi language speaker recognition using the language labels as the “channel” factor, and language recognition using the channel type (say, microphone type) as channel factor. The identity of the channel does not need to be known during scoring since the likelihood-ratio is computed by marginalizing over it.

Experiments using the proposed method for speaker recognition will soon start and will be reported on a separate publication.

References

- [1] S. Prince, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proceedings of the International Conference on Computer Vision*, 2007.
- [2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
- [3] A. Sizov, K. A. Lee, and T. Kinnunen, “Unifying probabilistic linear discriminant analysis variants in biometric authentication,” in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2014, pp. 464–475.
- [4] N. Brummer, “EM for probabilistic LDA,” Available at <https://sites.google.com/site/nikobrunner>, Tech. Rep., 2010.
- [5] J. Besag, “A candidate’s formula: A curious result in bayesian prediction,” *Biometrika*, vol. 76, no. 1, pp. 183–183, 1989.
- [6] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel, “Mixture of PLDA models in i-vector space for gender-independent speaker recognition.” in *Interspeech*, 2011, pp. 25–28.